# Organisation of a cliometric database

Pellier, Karine



**Empfohlene Zitierung / Suggested Citation:**
Pellier, K. (2005). Organisation of a cliometric database. *Historical Social Research*, *30*(3), 286-298. https://doi.org/10.12759/hsr.30.2005.3.286-298

# Organisation of a Cliometric Database

*Karine Pellier*[*]

**Abstract**: Often the empirical validation of theoretical assumptions starting from retrospective time series imposes the handling of an important volume of data. Once collected, this data can be structured in order to facilitate their conservation. This paper is particularly interesting in this intermediate phase and proposes to describe the organization of a database intended to be used as support for cliometric or econometric analysis. For that, we detail the stages of the creation of a database whose main objective is to store, organize and structure a unit of statistical series resulting from the satellite account of the Spanish education system. Our framework, even if it is based on an example, should be general for data management.

## Introduction

With the origin of the New Economic History, one finds the will of economists (Kuznets, 1941; Mitchell, 1927; Schumpeter, 1954) to exceed the weaknesses of a traditional approach too often limited to dissociate economic analysis of historical facts. Revealed in Conrad and Meyer (1957, 1958) works, it is popularized by Fogel and North, two economists historians honoured with the Nobel Prize of economy in 1993.

[*] Address all communications to: Karine Pellier, LAMETA, Université Montpellier I, Faculté des Sciences Economiques, Avenue de la Mer, C.S. 79606, 34960 Montpellier Cedex 2. France. E-mail: pellier@lameta.univ-montp1.fr.
The author would like to thank Nicolas Daures and Claude Diebolt for useful suggestions and comments.

The New Economic History is defined as a discipline of synthesis which judiciously combines economic theory and quantitative methods in the interest of history and economy. This *cliometric* renews the problems of the economic history and tries to provide, primarily on the basis of original statistical series, a reinterpretation of outstanding historical phases. The first works will concern more particularly United States growth at the XIXème century, the conditions of emergence of the industrial revolution or the great crises of modern history. Thereafter, with the renewal of the topics, growth theories will be the subject of a re-examination in order to propose answers to the great interrogations historic: what are the determining factors of a durable growth? What is the technical progress role, the human capital role? And so on[1].

Understanding the development or the decline of economies require resort to technics of modeling and to subject the assumptions which result to the test of history. So the work of the researcher is multi-field. He must have a sufficiently complete "toolbox". The instruments of analysis, borrowed from the statistics, must be powerful and suitable insofar as the results can lead to validate or refute theoretical approaches. The data to which will apply these quantitative methods are essential. One of the innovative aspects of the cliometric is the method which it used to collect and organize statistical data. The choice of the statistical variables is difficult because it must reflect as much as possible economic reality. In fact the economic models dictate this choice, but the data are not always directly or completely available. Often long term analyses require approximations or the rebuilding of the time series.

Databases constitute an interesting tool in preparing the data with the econometric treatments. Various arguments justify their implementation. They are able to manage significant volumes of data and to organize them in a logical way, i.e. without redundancies. They integrate rules of validity and consistency check during the seizure of information. It is also possible to share in network their content with made safe accesses, and so on.

This paper is particularly interesting in this phase of preparation of data and proposes to detail the organization of a database whose finality is double. Firstly, this one must file in an organized way economic and demographic data relating to Spain during XIXème and XXème century. The objective is that it becomes a support to carry out analyses of causality between the education system and the economic system of this country. On the other hand, this base will complete the education field of a more significant base: the database CAROLUS conceived and updated by N. Daures since 2001. This latter already contains data gathered in fields (education, demography, employment,...) and placed at the disposal of a team of researchers via a network.

Our framework, even if it is based on an example, should be general for data management.

---

[1] For a more complete cliometric presentation, the interested reader can visit the web site of the French Cliometric Association: http://www.cliometrie.org.

# 1. Implementation of the Base

The logical organisation of the database is based on the installation of the relational model of Codd (1970). First of all, this model consists in organizing and gathering in tables all the data having a common structure. These *entities* are then put in relation one to another to ensure the exploitation of the base as a whole.

This base is intended to store economic and demographic data relating to Spain over the period 1850-1965. The data, which appear in the form of times series, concern the total population of Spain, the total of the public expenditure of the State, the price index worked out by L. Prados (1993) and the national income at factors cost. One also has the education expenditure of the State, in current prices, distributed according to seven categories: expenditure as regards administration for teaching, expenditure for teaching primary education, secondary, technical, university, professional and of physically or mentally handicapped children. Concerning the author and the source of the collected data, they were compiled by C Diebolt (1999, 2000) and in majority come from the public finance directories published by the Spanish Ministry of economy and budget. The series were not modified nor corrected, however, they have missing data over the period of 1936 to 1940, period which correspond to the Spanish civil war. Further information complete these series: status of the data[2], measuring unit, references of publication and particular comments.

The simplicity of the organization of the base is guaranteed by a restricted number of tables. The sources and the units of the series being varied, it seems relevant to group these information in three tables. The first one, named "Series" gathers only the detail of the series: Title of series (full and abbreviated title), measuring unit, source, comments and finally the author of the series, i.e. the person who collected the data. The second table "Data" stores all the statistical data: year, value and status. The last one "Geographical areas" precise the space geographical of the data. It should be noted that this database contains only data on Spain, however, the database CAROLUS contains already data of other countries. The addition of this last table aims to facilitate the integration of our base in base CAROLUS.

The structure of a table is similar to that of a worksheet of a spreadsheet. The columns correspond to fields, i.e. elements which contain information in the same way standard and in the same nature. The data is stored in the form of records; a record is a line that contains several fields. For each one of these fields, it's advisable to specify the adapted type of data, its size, the indexings, etc. It's possible to record information of any kind in the fields thanks to a large variety of the existing types. Nevertheless, there are two main fields: those

[2] The status of the data informs us about its quality, i.e. if the data is that of the source, estimated, missing and so on.

containing numbers (numeric data) and those receiving text. Even if databases are particularly adapted to the bulky data management, it's preferable to minimize the volume of the base. A base of minimal size will be increasingly simpler to handle; faster calculations and faster transfers via Internet. Also, the choice of the type of data must be followed by a suitable value of the lengh of the field. It's significant to know with certainly the data type to be stored because any later modification of the type or the size of the fields is likely to generate problems, especially on the level of the posting of the data. According to these recommendations, we associate to each field of the tables "Series", "Data" and "Geographical Areas" an adequate data type as well as the lengh which corresponds (fig. 1). The default value of a text type is of 50 characters but it may be adjusted according to the data. Moreover, *long integer* is used for numbers without decimals or automatic numbers and *simple* is used for decimal numbers.

Fig. 1: Properties of the tables

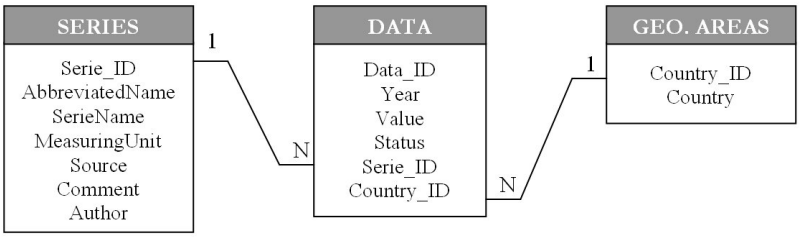
| Table | Field | Data Type | Field size | Indexed |
|---|---|---|---|---|
| SERIES | Serie_ID | Auto. number | Long integer | Yes without repetition |
| | SerieName | Text | 250 | No |
| | Abbreviated-Name | Text | 10 | No |
| | MeasuringUnit | Text | 50 | No |
| | Author | Text | 50 | No |
| | Source | Memo | 65 535 | No |
| | Comment | Memo | 65 535 | No |
| DATA | Data_ID | Auto. number | Long integer | Yes without repetition |
| | Year | Numeric | Integer | No |
| | Value | Numeric | Double | No |
| | Status | Numeric | Byte | No |
| | Serie_ID | Numeric | Long integer | Yes with repetition |
| | Country_ID | Text | 10 | Yes with repetition |
| GEO. AREAS | Country_ID | Text | 10 | Yes without repetition |
| | Country | Text | 50 | No |

Each table gathers all the data with the same logic design and each one of them has an identifying field or *primary key*. A primary key uniquely identifies each record in a table. Thus, for the table "Series", the identifier "Serie_ID"

identifies in a single way each series (population, price index …). By convention, the primary keys have got neither null values nor repeating values. The field "Value" refers to the statistical data itself and the field "Status" gives us an indication about the nature of the data (0: missing, 1: valid, 2: estimated, 3: calculated).

A relationship is made between two tables by matching the values of a foreign key of one table with the values of the primary key in another (fig. 2). A foreign key is a field whose values are the same as the primary key of another table. Primary and foreign keys are fundamental because they enable tables in the database to be related with each other. The field "Serie_ID" of the table "Data" is a foreign key which returns to the statistical series to which the data belongs. In the same way, the field "Country_ID" indicates the country of reference and returns to the table "Geographical Areas". A precaution must be taken during the connection of the tables. The dependent fields must be imperatively in the same type and of the same lengh. But if the primary key is an automatic number then the external key will be necessarily a numeric.

Fig. 2: Relational diagram of the base




The relationship between the tables is of type one-to-many (1-N). It means that the field of the table containing the primary key is in relation to one or more records of the table containing the external key. For the tables "Series" and "Data", that implies that a serie is in relation to many data (the data are annual and the period is spread out of 1850 to 1965).

The interest of the relationships is to improve the access to the tables. They will make it possible to exploit as a whole all the data of the base. However, they must check rules and in particular the referential integrity of the data. The role of this latter is to ensure the validity of the seizures. It constitutes a safety in that it limits the errors of external keys and prohibits any suppression or inopportune modification of primary keys or records. Concretely, for our base, it will be possible to complete the table "Data" only by records corresponding to a record of the tables "Series" and "Geographical Areas". The referential integrity rule makes sure that every foreign value match a primary key value in an associated table.

The relational model that we've just described prepares the establishment of the data in a database management system (DBMS). A DBMS-R is software based on the relational model and is intended to manage large quantities of information, persistent, reliable and shareable between several users. The structure of our base being not very complex, we choose its implementation with the Access software[3].

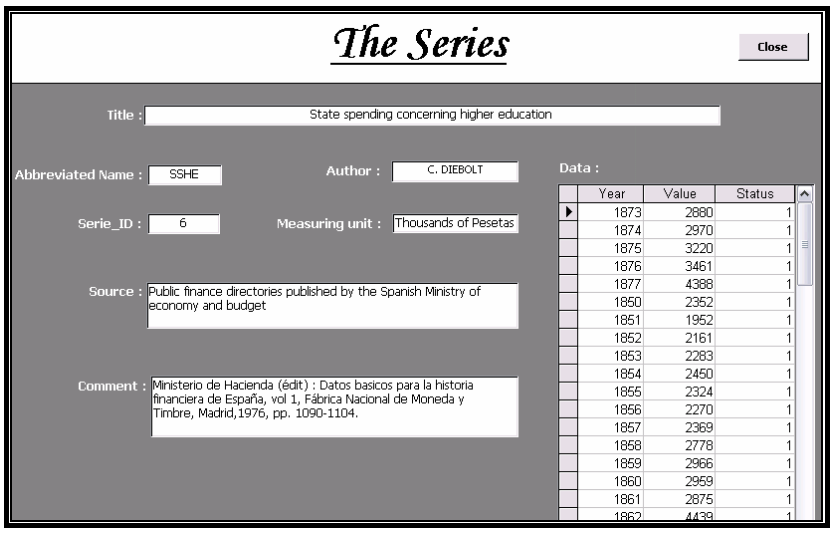
## 2. Transfer and Consultation of the Data

At the end of the installation of the structure of the base and tables, we have to import or enter the information in the first place. The data being stored in traditional supports (worksheets), their transfer in the table „Data" is carried out by a procedure of importation. The use of the spreadsheet for the manual seizure of the data is convenient because the functions can detect anomalies (missing values, inconsistencies…). Tables "Series" and "Geographical Areas" are informed manually either directly, or by the intermediary of a form. The forms like the tables, the queries and the reports are the major components of Access which allow the handling of the data. A form is a window to consult, modify or add data resulting from one table or query. It makes easier the handling of the base and consequently limits the errors of seizures of the user.

For this database, it seems interesting to create two forms. The first is based mainly on the table "Series" but also depends on the table "Data". This form allows us to consult in the same window and at the same time all the information relating to the selected serie and the associated statistical data (fig. 3). For safety measures, the modifications such as the suppression of data are forbidden. But the interest of a form is also to be able to associate to certain buttons of the form a particular action. The action started by a clic of the mouse is a macro i.e. a procedure written in a specific language: the VBA (Visual Basic Application). When a form is created, it's easy to insert buttons which will be used to open another form, table, query, report or close the active form.

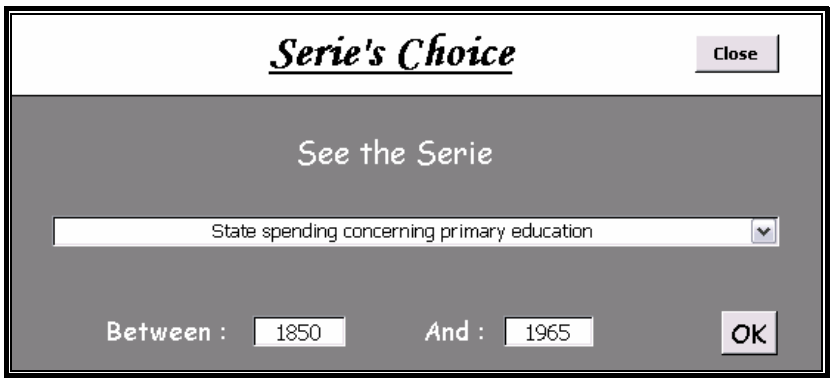
[3] The DBMS Oracle or SQL-Server, more powerful, are able to manage bases of very complex structure but on the other hand require more data-processing competences.

Fig. 3: Form of seizure/consultation of information




The period of complete study extends from 1850 to 1965. But it happens that the economist user of this base wishes to carry out an analysis over one period more restricted. In this case, we propose a second form “Serie’s Choice” allowing him to select the serie he wishes to see as well as the desired period (fig. 4). The default values define the limits of the interval. The characteristic of this form is to be independent with the tables. However, it is associated to a query which is based on the tables. The fields of this form will make it possible to specify the criteria of a parameter query intended to search for the data.

Fig. 4: Form of filtering of the data



# 3. Exploitation of the Base

The principal function of a database collecting an important volume of data is to retrieve particular information in order to analyze them. The queries are the most powerful tool to fulfill this function. They filter and view the researched records in a window similar to a worksheet. The types of queries are varied: select query to search data with conditions, parameter query, crosstab query, etc.

There is a dynamic interaction between tables and queries. When one modifies or adds data in a table, the query based on the fields of this table is updated automatically during its run. In order to avoid congestion, it's only the structure of the query that is recorded and not the datasheet. Nevertheless, every modification carried out on these sheets is reflected on the data of the corresponding table.

The procedure to create a query initially consists in defining the objective of the query, i.e. the data we are looking for. Then, it's advisable to identify the tables containing these data. Finally, it is necessary to define the criteria and the operations applied to the data. A query can combine data from multiple tables or queries on condition to having established as a preliminary the relationships between these objects.

We create for this base four queries with particular goals. The first one is a query by form. The form "Serie's Choice" is a custom form that prompts for a query's parameters. Once established, the structure of the query won't be modified. Indeed, in order to ensure the correct working of the base, the user is simply satisfied to enter the parameters of the form and doesn't have any need in any case to make modifications to the query. From this form, the user chooses the period and the serie he wishes to look. In the query, the criteria applied to the field "Year" define precisely the limits of the interval (fig. 5). For each one of these limits, it's necessary to indicate the name of the form and the name of the control which is associated. Control entitled "B" (respectively "E") of the form "Serie's Choice" is the lower limit (respectively higher) of the interval. These limits can contain only existing values in the field "Year" of the table "Data". Control "S" is a combo box which enables the user to choose the title of the serie which interests him. After the run of the query, the user views the filtered data on the datasheet.

Fig. 5: Criteria of a Parameter Query

| Field | Table | Visible | Criteria |
|---|---|---|---|
| Year | Data | Yes | Between [FORMS]![Serie's Choice]![B] And [FORMS]![Serie's Choice]![E] |
| Value | Data | Yes | |
| Serie_ID | Data | No | [FORMS]![Serie's Choice]![S] |

The second query will firstly extract the data relating to education spending at current prices and also calculate the annual sum of this expenditure. Given that it's a question of connecting several fields of the table "Data" we develop a particular request: "Crosstab Query". In order to visualize in a form of a table the annual value of each type of expenditure, the field "Years" is put on the line header and the field "AbbreviatedName" in the columns header (fig. 6). The logical operator *OR* is used to specify the criteria of the field "Abbreviated-Name". Its function is to restrict the viewed series only with those which interest us. The structure of this query is finished by the addition of a calculated field "Total" based on "Value". This field makes the sum on line, i.e. for each year, of the various types of spending.

Fig. 6: Criteria of a Crosstab Query

| Field | Table | Total | Crosstab | Ctiteria |
|---|---|---|---|---|
| Year | Data | Group by | Row heading | |
| Abbreviated-Name | Series | Group by | Column heading | |
| Value | Data | Sum | Value | |
| Total : Value | Data | Sum | Row heading | |
| Abbreviated-Name | Series | Where | | "SSAE" Or "SSPE" Or "SSPRE" Or "SSSE" Or "SSTE" Or "SSHE" Or "SSHAE" |

An effective database should not receive new serie which could be obtained by combining the data already included. Let us admit that the economist wishes to know the share of the public expenditure devoted to education over the entire period. It's useless to add a new serie to the base since one has at the same time the total of the public and the educational spending, obtained by the crosstab query[4]. Thus, this new query is based on the fields "Year" and "Total" of the crosstab query and on the fields "Year", "Serie_ID" and "Value" of the table "Data" (fig. 7). So to ensure the correct working of the query, the two same fields "Year" are connected by an equijoin. One adds a calculated field, the field "Ratio" which will carry out the ratio between the "Total" field (education spending) and the "Value" field (public spending). The datasheet viewed only two fields: year and value of the ratio for each year.

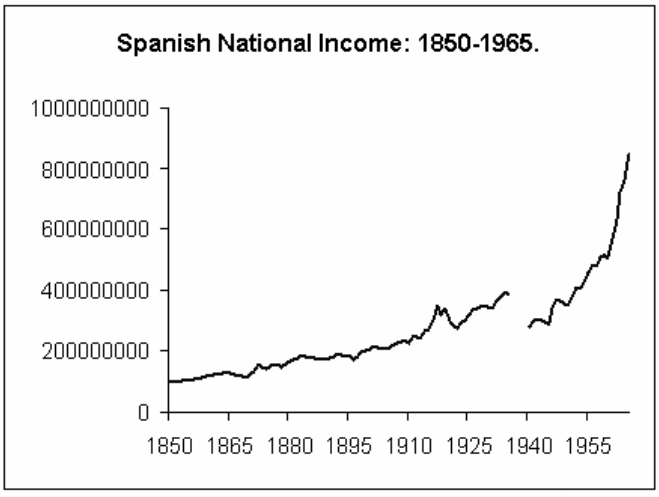
[4] In the same way, education spending at constant price can be obtained starting from the series of the price index.

Fig.7: Criteria of a Select Query

| Field | Table/Query | Visible | Criteria |
|---|---|---|---|
| Year | Education spending | Yes | |
| Total | Education spending | No | |
| Year | Data | No | |
| Serie_ID | Data | No | 15 |
| Value | Data | No | |
| Ratio: [Total]/[Value] | | Yes | |

The finality of the last query is to retrieve the serie of the Spanish national income. Once extracted, this serie will be graphically represented over the initial period. This last query is based on the fields "Year", "Value" and "Serie_ID" of the table "Data". As one wishes to post the values of a particular serie, it just takes to indicate in the criterion zone of the field "Serie_ID" its number. Consequently we seize number 17 for the serie of the Spain's national income. From this query, a report is created in order to obtain a graphic representation of the serie (fig. 8). Reports provide the capability to quickly produce formatted summaries of the data contained in one or more tables or queries. They include graphics or calculated fields. Thanks to the graphic report related to query, the user can visualize one or more series simultaneously.

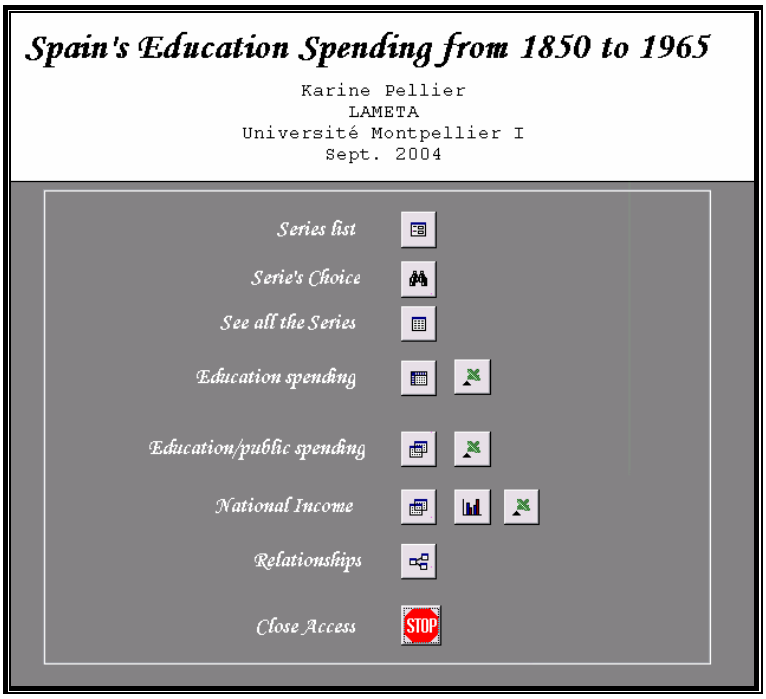
Fig. 8: Graphic report based on a query



The queries are interesting objects to question a base, however when one wishes to carry out more complex calculations, it's rather recommended to export the series towards other more adapted software (spreadsheet…). For this base, several buttons with macros make it possible to export and thus to convert with the format chosen (here Excel) the data. This data could be imported thereafter in other scientific software, such as the E-Views software, in order to be subjected to statistical processing. The spreadsheet seems to be a good complement to the DBMS. It happens previous to the data acquisition and after to facilitate their transfer towards specific software.

We'll finish this database by the overview of a general menu (fig. 9). This form, which is viewed since the opening of the base, aims to give general indications on the database and also to facilitate its use by the user. It gathers, in only one place, all of the buttons giving access to the various objects created for the exploitation of the base. Erroneous handlings of the base are limited owing to the fact that these various objects open on a reader mode alone.

Fig. 9: General menu



# Conclusion

This database finished is now operational to be exploited. It answers the starting requirements: organizing data in order to prepare them with future statistical processing and in particular with tests of causality. Its simple and concise structure guarantees its best working. The export of the data structured and filtered towards other software or the connection of the tables to database CAROLUS can be carried out without any difficulty. As for the users having a minimum of data-processing competences, they will easily be able to add the base with new objects that answer they need or to simply update it.

Databases are very powerful tools to organize data in a flexible manner. They become truly an essential stage to any cliometric analysis based on multiple retrospective series. Whatever the nature of the data may be, the design of a base always imposes the same step. A preliminary work of reflexion is to be carried out before any handling of the data. It aims at identifying the user of the base and the use he wishes to make. The objectives of the base are thus determined by raising the questions to which it must bring an answer. Starting from there, the structure of the base can be implementing: tables are created to store the data; forms to consult or update the data; queries to retrieve the data and reports to print the results. All these operations are effected by the intermediary of a relatively convivial data-processing interface.

# References


BOUZEGHOUB M. (1998). *Le Modèle relationnel : Algèbre, langages, applications*, Collection Les bases de données en question, Hermès.

CODD, E. (1970). "A Relational Model of Data for Large Shared Data Banks", *Communications of the ACM,* vol. 13, n°6, pp 377-387.

CONRAD A. (1957). "Economic Theory, Statistical Inference and Economic History", *Journal of Economic History*, vol. 17, n°4, pp. 524-544.

CONRAD A. & MEYER J. (1958). "The Economics of Slavery in the Ante Bellum South", *Journal of Political Economy*, vol. 66, n°2, pp. 95-130.

DAURES N. (2004). „La base de données CAROLUS. Données économiques et démographiques sur l'éducation en France aux XIXème et XXème siècle.", *First Workshop on Cliometrics & Econometric History*, AFC, CEROM de l'Ecole Supérieure de Commerce de Montpellier & UMR LAMETA, 10 septembre 2004.

DIEBOLT C. (1999). "Gouvernment Expenditure on Education and Economic Cycles in the Nineteenth and Twentieth Centuries. The case of Spain with special Reference to France and Germany", *Historical Social Research*, vol. 24, n°1, pp. 3-31.

DIEBOLT C. (2000). *Dépenses d'éducation et cycles économiques en Espagne aux XIXème et XXème siècles,* L'Harmattan, Paris.

DIEBOLT C. (2004). „La cliométrie se rebiffe!“, *First Workshop on Cliometrics & Econometric History,* AFC, CEROM de l'Ecole Supérieure de Commerce de Montpellier & UMR LAMETA, 10 septembre 2004.
FOGEL, R. (1964). *Railroads and American Economic Growth: Essays in Econometric History,* The Johns Hopkins University Press, Baltimore.
FOGEL, R. (1965). "The Reunification of Economic History with Economic Theory", *American Economic Review*, vol. 55, n°2, pp. 92-98.
FOGEL, R. (1994). "Economic Growth, Population Theory, and Physiology: The Bearing of Long-Term Processes on the Making of Economic Policy", *American Economic Review*, vol. 84, n°3, pp. 369-395.
KUZNETS S. (1941). "Statistics and Economic History", *Journal of Economic History*, vol 1, pp 26-41.
MCCLOSKEY, D. (1976). "Does the Past Have Useful Economics ?", *Journal of Economic Literature*, vol. 14, n°2, pp. 434-461.
MCCLOSKEY, D. (1987). *Econometric History*, Macmillan, London, 1987.
MINISTERIO DE HACIENDA (ED.). (1976). Datos basicos para la historia financiera de España (1850-1975), 2 vols., Fabrica Nacional de Moneda y Timbre, Madrid.
MITCHELL W.C. (1927). *Business Cycles: The Problem and its Setting,* National Bureau of Economic Research, New York.
NORTH D. (1994). "Economic Performance Through Time", *American Economic Review*, vol. 84, pp. 359-368.
PRADOS DE LA ESCOSURA, L. (1993). "Spain's Grows Domestic Product, 1850-1990: A New Series", *Working Paper*, D-93002, Universidad Carlos III de Madrid, March.
ROLLINAT R. (1997). *La nouvelle histoire économique,* Editions Liris, Collection Perspectives Contemporaines, Paris.
SCHUMPETER J. (1954). *Histoire de l'analyse économique*, Traduction française, Gallimard, Paris, 1983.
Website of The French Cliometric Association (AFC): http://www.cliometrie.org
VIESCAS J. (2001). Microsoft Access version 2002 au quotidien, Microsoft Press.